Vol. 176: 303–306, 1999 | MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser | Published January 18

COMMENT

How many statistical tests are too many? The problem of conducting multiple ecological inferences revisited

Pedro R. Peres-Neto*

Department of Zoology, University of Toronto, Toronto, Ontario M5S 3G5, Canada

In several situations ecologists are interested in addressing multiple statistical tests among samples (e.g. a series of species or abiotic/biotic variables). The most common application is to carry out all 2 by 2 comparisons between all samples or to perform only those comparisons of interest. In both cases, differences between samples are subjected to a statistical test that provides the significance of the contrast. The most intuitive approach would be to test the significance of each comparison separately. For example, a community ecologist interested in comparing differences between microhabitat use by several species does pairwise chi-squared tests between them. For each pair, the null hypothesis that 2 species in particular have similar habitat preferences is rejected or not, based on the significance of the chi-squared value between the 2 species. Nevertheless, as I will show here, the approach of performing multiple statistical tests independently is inadvisable.

The significance level, or alpha value, established *a priori* is the probability of committing the so-called Type I error, which is the sampling frequency at which the null hypothesis will be rejected when it is true. In other words, if a significance level of 0.05 is chosen, for 100 sample values of the test being conducted, 5 of them will be considered significant when in reality (i.e. in the population) they are not. Choosing the appropriate significance level allows the researcher to control unusual random differences in contrast to significant differences. Nevertheless, this probability is related only to a single test and cannot be maintained when multiple tests are being conducted. For instance, suppose that a researcher wants to conduct pairwise comparisons using correlation values between 10 variables with a significance level of 0.05. Thus, it should be expected that at least 2.25 (45 times 0.05) significant differences should be found purely due to random correlations. Here, if only 2 significant contrasts with alpha set at 0.05 had been encountered, these results should be interpreted with caution. Consequently, as the number of tests being conducted increases, more 'significant' values are found. To show this, I simulated multiple correlation contrasts ranging between 3 and 60 variables (i.e. 3 to 1770 comparisons). For each multiple contrast, I performed 10 replicates, averaging the number of significant results, using random normally distributed variables with 60 observations. Using a level of significance of 0.05, a strong linear relationship between number of contrasts and significant results due only to random correlations is observed (Fig. 1). The slope of the relationship is $b = 0.051$, demonstrating that the number of 'significant' values due to random correlations is actually 5% (0.05) for all comparisons.

Zar (1996) has warned biostatisticians that 'two-sample tests, it must be emphasized, cannot be utilized validly to test multisample hypotheses'. Rice (1989) cautioned researchers in evolutionary biology of the danger of interpreting significant results of 2-sample tests. Although such issues have been raised in other fields for 2-sample (or variable) tests, as I will show later, the problem of multiple inferences is not only related to pairwise comparisons using tests such as *t*-tests and Mann-Whitney tests, Pearson and Spearman correlations and the chi-squared statistic, but to any circumstance where multiple tests are being conducted. Even though on many occasions ecologists have used the correct approach, more rigorous applications are still required. Thus, my goal is to call the attention of ecologists to the problems related to multiple inferences in general, not only pairwise comparisons, showing some solutions.

The main concern here is to control the number of inflated 'significant' values with the number of tests

*E-mail: pperes@zoo.utoronto.ca

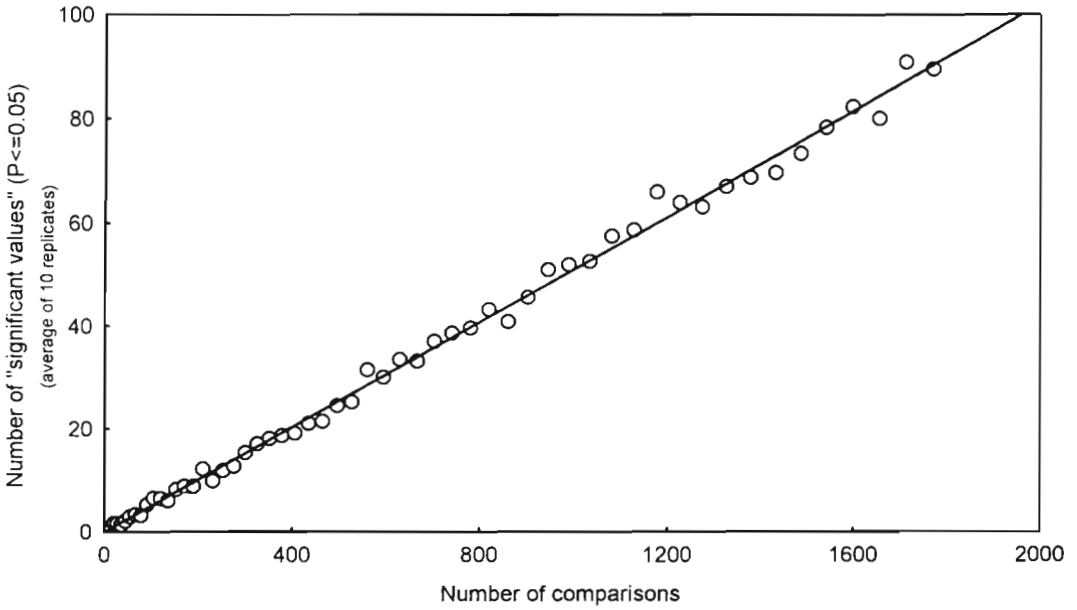

Fig. 1. Correlation between the number of contrasts in multiple comparisons and the number of inflated 'significant' values

being conducted just because of chance (Fig. 1). To avoid this problem (i.e. Type I errors), the probability of rejecting the null hypothesis should be adjusted to account for the number of tests being conducted. Using a Bonferroni inequality, it has been shown that to ensure an overall significance level equal to or smaller than alpha for all k desired tests, a significance of alpha/k should be used (Miller 1981). For this reason, the most simple and straightforward method for multiple inference is the Bonferroni test. Let us say that 10 samples are being compared (i.e. k = 45 comparisons) with an alpha = 0.05, the new significance level for all pairwise comparisons should be $0.05/[(10 \times 9)/2] = 0.001$. Unfortunately, by using such a small alpha, the acceptance range becomes too wide and a large number of Type II errors (i.e. accepting the null hypothesis when it is not true) would occur and the tests would have limited statistical power. Thus the Bonferroni test should not be used with a large number of comparisons. To circumvent this problem while still preventing Type I error, the confidence intervals for the statistic being applied should be modified to account for simultaneous inferences.

Several tests for performing multiple comparisons exist. The differences between them are mainly due to parameter requirements (i.e. normality and homoscedasticity), criteria for how Type I errors will be defined and the power of the test. Classical approaches include methods such as Tukey, Scheffé, Least Significant Difference, Duncan, and Newman-Keuls tests (Klockars & Sax 1986, Toothaker 1991). Although some of these tests have been used to compare other statistics, such as correlation indexes (see application of Tukey test for the Pearson correlation index, Zar 1996), they are usually restricted to comparisons of differences between means. Thus, these methods often cannot be readily applied to statistics like chi-squared values, overlap indexes, coefficients of variation, or any other statistics of interest. In addition, statistics for which the probability distribution is unknown and some Monte Carlo procedure (Manly 1997) is used for estimating the probability of rejecting the null hypothesis are not subject to standard techniques for multiple inferences. For these reasons, I decided to present an application of the sequential Bonferroni method developed by Holm (1979) which can be applied to almost any statistic because of its non-parametric nature. This technique was developed to increase power over the simplest case of Bonferroni correction that, as discussed above, is very conservative in rejecting null hypotheses. Although improvements about gaining power (i.e. rejecting the null hypothesis when it is not true), at least when applying it for comparing means, over the sequential Bonferroni technique have been presented (e.g. Schaffer 1979, 1986, Holland & Copenhaver 1988, Hsiung & Olejnik 1994, Seaman 1997), Holms's (1979) approach remains the simplest (Holland & Copenhaver 1988) general and easily applicable method.

To illustrate the sequential Bonferroni approach, I will consider an example applied to Pearson correlations between 5 morphological characteristics for 47 fish species of eastern Brazilian fishes (Peres-Neto unpubl. data) (Table 1). Considering an alpha of 0.05, 5 correlations are significant. The application of a se-

Table 1. Pearson correlations and associated raw probability values (in parentheses) between 5 morphological characteristics for 47 species of eastern Brazilian fishes

Characteristic	1	2	3	4	5
1	1.000				
2	0.110 (0.460)	1.000			
3	0.325 (0.026)	0.345 (0.018)	1.000		
4	0.266 (0.070)	0.130 (0.385)	0.142 (0.340)	1.000	
5	0.446 (0.002)	0.192 (0.196)	0.294 (0.045)	0.439 (0.002)	1.000

Table 2. Application of the sequential Bonferroni correction for correlation values in Table 1. *Significant values at alpha = 0.05 before and after correction

i	Correlation	Probability before correction	alpha/ $(1 + k - i)$	Significant values after correction
1	0.439	0.002*	0.005	*
2	0.446	0.002*	0.006	*
3	0.345	0.018*	0.006	
4	0.325	0.026*	0.007	
5	0.294	0.045*	0.008	
6	0.266	0.070	0.010	
7	0.192	0.196	0.013	
8	0.142	0.340	0.013	
9	0.130	0.385	0.025	
10	0.110	0.460	0.050	

quential Bonferroni correction involves 3 steps. (1) Calculate the exact probability of having a random value equal or higher than the observed value for the statistic of interest. (2) Sort in ascending order the k comparisons probability values, referring to them as p_i (p_1, p_2, p_3, .., p_k). Tied p values can be ordered arbitrarily. (3) Compare each p value with the following inequality: $p_i \leq \text{alpha}/(1 + k - i)$. If the inequality is not met, i.e. $p_i > \text{alpha}/(1 + k - i)$, then consider the correlation not significant at alpha = 0.05. Since the p values are ordered, the comparisons can be stopped after encountering the first inequality that is not met because all the subsequent correlations are not significant. Note that after the correction only 2 correlations remained significant ($r_{15} = 0.446$ and $r_{45} = 0.439$) (Table 2).

There is a common misunderstanding that the problem of multiple inferences is only related to pairwise contrasts and in extreme cases only to parametric statistics. An important element to keep in mind is that, whenever several statistical tests are being performed, rejection due to chance is highly related to the number of tests being conducted just because of the law of large numbers (Miller 1981). In any case, regardless of the number of tests, corrections should be always conducted. Some additional kinds of inappropriate approaches of multiple inferences in the ecological literature include: comparing several simple and multivariate linear regressions for the same variables applied to different samples (e.g. weight-length for different species, density-body size and species-area relationships for different regions); several analyses of variance applied to different variables; several analyses of covariance applied to several different samples (e.g. species or sites). Some extremes examples in the ecological literature include papers where both correct (e.g. when comparing sample means) and incorrect (e.g. comparing correlations) approaches were applied.

A final consideration relates to the number of tests to be conducted. The same problem of fixing the significance level before samples are taken applies to the number of contrasts. It is inadvisable to change the number of tests after the sampling has been completed (i.e. conduct exploratory analyses), as this can change the control of the significance level over random differences. The following example, in which a researcher has decided to investigate the differences among 6 sample means, illustrates this point. For simplicity, let us assume that the null hypothesis is true. Thus, the distribution of differences among means of all possible samples would be near zero, and any large difference would be due to chance. Assume now that the researcher has decided to compare for whatever reasons only the largest sample means. This deliberate (i.e. not random) choice can influence the rejection of the null hypothesis, because under this new scenario this 'chosen' difference will always exceed the differences expected by random. Since the significance level does not control for systematic errors, it will not hold in this case. On the other hand, if one decides to decrease the number of samples, the significance level should be kept the same as for the original number of comparisons planned. For the sequential Bonferroni test, all comparisons should be performed as planned. If one decides to increase the number of comparisons, the probability of committing a Type I error will increase as well (Fig. 1), but one can control for this by using more conservative tests (e.g. Scheffé test). Unfortunately, again, more conservative methods might reduce the power of rejecting the null hypothesis when it is false. For comparing means, the standard Bonferroni test has been shown to be more powerful than Scheffé's method when the number of comparisons is no more then 47 (Milliken & Johnson 1992).

Multiple comparisons of sample means have been the subject of many more studies than any other statistical application in both the ecological (e.g. Day &

Quinn 1989) and the statistical literature (e.g. Carmer & Swanson 1973, Jaccard et al. 1984, Hsiung & Olejnik 1994). Nevertheless, biologists deal with a broad range of situations, in which different kinds of measurements are necessary. The sequential Bonferroni method is a suitable solution for most of them, but its behavior in different situations is not fully realized, and studies of this issue should be developed. My goal in this comment is to convince ecologists to use caution when conducting multiple inferences, and that more effort in applying the right tools should be given.

Acknowledgements. I thank Donald A. Jackson, John R. Stone and Wade B. Worthen for their comments on the manuscript. Funding was provided by a CNPq Doctoral Fellowship.

LITERATURE CITED

Carmer SG, Swanson MR (1973) An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. J Am Stat Assoc 68:66–74

Day RW, Quinn GP (1989) Comparisons of treatments after an analysis of variance in ecology. Ecol Monogr 59:433–463

Holland BS, Copenhaver MD (1988) Improved Bonferroni-type multiple testing procedures. Psychol Bull 104: 145–149

Holm S (1979) A simple sequential rejective multiple test procedure. Scand J Stat 6:65–70

Hsiung T, Olejnik S (1994) Power of pairwise multiple comparisons in the unequal variance case. Communication in Statistics - Simulation and Computation 15:691–710

Jaccard J, Becker, MA, Wood G (1984) Pairwise multiple comparison procedures: a review. Psychol Bull 96:589–596

Klockars AJ, Sax G (1986) Multiple comparisons. Sage Publications, Beverly Hills

Manly BFJ (1997) Randomization, bootstrap and Monte Carlo methods in biology. Chapman and Hall, New York

Miller RG (1981) Simultaneous statistical inference. McGraw Hill, New York

Milliken GA, Johnson DE (1992) Analysis of messy data. Chapman and Hall, New York

Rice WR (1989) Analysing tables of statistical tests. Evolution 43:223–225

Schaffer JP (1979) Comparison of means: an F followed by a modified range procedure. J Educ Stat 4:14–23

Schaffer JP (1986) Modified sequentially rejective multiple test procedures. J Am Stat Assoc 81:826–831

Seaman MA (1997) Tables for pairwise multiple comparisons using Shaffer's modified sequentially-rejective procedure. Communication in Statistics - Simulation and Computation 26:687–705

Toothaker LE (1991) Multiple comparisons for researches. Sage Publications, Beverly Hills

Zar JH (1996) Biostatistical analysis. Prentice-Hall, Englewood Cliffs, NJ